HW3 report

姓名: 王國豪　　學號: b07901032

1. Make

首先是組建助教提供的 srilm 時遇到了一些問題，決定使用官方 1.7.3 的 release。再者，在 link 時會遇到以下 undefined reference to `xxx`錯誤:

```
1|  /root/srilm/misc/src/tls.cc:15: undefined reference to `pthread_key_create'
```

在 makefile 中 link 時加入以下綠字選項可以解除這個現象，若是有更多這類錯誤，可能要加上更多的 flag 讓 compiler 可以找到目標:

```
1|$(TARGET): $(OBJ) -loolm -ldstruct -lmisc
2|     @$(CXX) -lpthread -fopenmp -lm -lz -o $@ $^
```

2. Mapping

決定使用 c++實作就是一連串折騰的開端。我為注音與文字各開了一個 class，因為注音要按照順序排列，因此需要 overload operator <來讓 sort 可以運作。觀察 big5 table 涂以為只要將第一個 char 減去 ”NUL” 乘上 256 加第二個字元就可以比較，結果卻產生從 ㄐ 開始的序列，最後是直接使用第二個字元做比較。

3. Mydisambig

實作上使用 map<string, set<string> >來存放 zhuyin-big5.map，使用 vector<pair<vector<string>,float> >構築整個 viterbi 的 $\delta_i(t)$並存放最佳路徑。基本上就是照著 viterbi 的演算法將每一行讀進來並填上機率上較接近的結果。

4. Srilm/disambig vs Mydisambig

A. Performance

	srilm(trigram)		my(bigram)	
	time(s)	mem(Mb)	time(s)	mem(Mb)
test_data#1	6.15	59.88	6.17	22.15
test_data#2	7.65	60.89	12.03	25.15
test_data#3	5.08	59.74	7.57	22.35
test_data#4	5.4	60.28	9.7	23.09
test_data#5	4.64	59.99	7.51	22.64

B. Observation & Accuracy

首先是效能的部分，因為 Srilm 的功能繁複多樣，因此多用許多記憶體是非常合理的事情。時間上也是 srilm 較為快速，應該是有針對演算法做優化的部分。

再來觀察解碼的文檔:

原檔:

```
1| <s> 厂 視 丁 聞 開 礻 一 喜 李 四 端 金 素 梅 明 搭 檔 雙 主 播 </s>
2| <s> 華 尸 丁 聞 將 在 冂 天 年 第 一 天 推 礻 凵 旦 雙 主 播 </s>
3| <s> 夕 外 丬 極 合 卩 </s>
4| <s> 在 會 虫 夕 事 郭 力 昕 提 弓 討 分 與 台 尸 ㄎ 略 聯 冂 的 可 丁 性 </s>
```

Mydisambig:

```
1| <s> 忽 視 新 聞 開 場 迎 喜 李 四 端 金 素 梅 明 搭 檔 雙 主 播 </s>
2| <s> 華 社 新 聞 將 在 明 天 年 第 一 天 推 出 約 旦 雙 主 播 </s>
3| <s> 對 外 積 極 合 作 </s>
4| <s> 在 會 長 董 事 郭 力 昕 提 案 討 論 與 台 商 策 略 聯 盟 的 可 行 性 </s>
```

Srilm:

```
1| <s> 華 視 新 聞 開 朝 野 喜 李 四 端 金 素 梅 明 搭 檔 雙 主 播 </s>
2| <s> 華 視 新 聞 將 在 明 天 年 第 一 天 推 出 元 旦 雙 主 播 </s>
3| <s> 但 外 界 極 合 作 </s>
4| <s> 在 會 中 毒 事 郭 力 昕 提 案 討 論 與 台 商 策 略 聯 盟 的 可 行 性 </s>
```

可以發現 bigram 和 trigram 在中文文本裡的行為非常的明顯，例如「華視新聞」這樣四個字長的 trigram 可以完美地找到正確的字，但在像是「會中/董事」這樣兩組比較不相干的詞就會因為 trigram 的行為導致最後四個字合在一起不正確。可惜沒有時間多完成輸出文本精準度的分析，無法驗證心中一些對 trigram 與 bigram 行為的理解。